Anticipating Failures: What Should Predictions Be About?

Erik Hollnagel
Professor, Ph.D.
Graduate School for Human-Machine Interaction
University of Linköping, SE-581 83 Linköping
Sweden
eriho@ikp.liu.se; eriho@ida.liu.se

Summary: Accident analysis and performance predictions have traditionally been pursued in separate ways, using different concepts and methods. This has made it difficult to use the experiences from accident analysis in performance prediction. As a result, performance prediction is still focused on the concept of individual "errors", despite overwhelming evidence that accidents are caused by a concatenation of conditions rather than a single action failure. It is argued that the anticipation of failures should be based on better models of how performance conditions determine actions, and that the inherent variability – or unreliability – of human performance is the noise rather than the signal.

1. INTRODUCTION

Accident analysis and performance prediction for human-machine systems have traditionally been pursued as two separate activities, despite the obvious fact that they refer to the same reality – namely the occurrence of unexpected events leading to unwanted outcomes. Accident analysis has been concerned about unravelling the complex of causes that might explain what happened, and preferably finding one or a few causes that could be considered the root or origin of the accident. Performance prediction has been concerned with trying to identify in advance the risks inherent in a system, in order to be able to change or modify the design so that these risk can be reduced or eliminated. In both cases a common motivation has been the dramatic rise since the 1970s in the number of cases where the causes of accidents have been attributed to incorrectly performed human actions. Although this does not by itself mean that there have been more "human errors", it expresses a distinct change in attitude towards the analysis of accidents and the commonly accepted set of causes (cf. Hollnagel, 1993a).

Accident analysis for systems involving human-machine interaction has always had a strong psychological flavour, looking toward "human error mechanisms" and various deficiencies of information processing that are supposed to occur in the human mind (e.g. Senders & Moray, 1991). In contrast to that, performance prediction has been dominated by the engineering quest for quantification, as epitomised by the PSA event tree, and models and methods have been constrained by that (e.g. Dougherty & Fragola, 1988). In both cases there has been a strong predilection for considering "human error" as a category by itself, referring either to complex models of how information processing can go wrong or to estimates of single "human error probabilities". This view persists despite a growing realisation that it is a gross oversimplification which fails to recognise the complexity and significance of human performance failures (Hollnagel, 1993a; Woods et al., 1994).

2. APPROACHES TO ACCIDENT ANALYSIS

The analysis of an accident is always based on an accident model, i.e., a conceptualisation of the nature of accidents, specifically how a set of causes and conditions may lead to an accident. Current accident models must account for the complex interaction between humans, technology, and organisations. The accident model may be explicitly formulated but is more often implicit, hidden in the assumptions that investigators make. Every accident model is based on the principle of causality, which states that there must be a cause for any observed event, and the models serve as guidance for finding the acceptable causes. In the following I will briefly consider the major changes to accident models since the 1950s, since these reflects the developments in the commonly agreed understanding of the nature of an accident.

2.1 Simple Accident Model

The first accident models tended to see accidents as caused either by failures of the technology or incorrect human actions, cf. Figure 1. Before the accident the system was assumed to be in a normal state, and an incorrect human action was seen as the primary cause of the accident. Accident classifications typically used the "human error" category as a kind of catchall, or garbage can, for accidents that could not be attributed to the failure of a technical component. The simple accident model corresponds to methods such as root cause analysis (Park, 1987; Cojazzi, 1993; Cojazzi & Pinola, 1994), which from a psychological view are relatively unsophisticated. In relation to the specific issue of human failures, the simple accident model is closely associated to the information processing point of view, which harbours three basic assumptions. Firstly, that there are reliable criteria of validity against which it is possible to measure a deviant response. Secondly, that psychological factors affect information processing and act to bias responses away from the standards considered appropriate. And finally that the human information processing system comprises a diverse range of limitations that are invoked under particular information processing conditions.

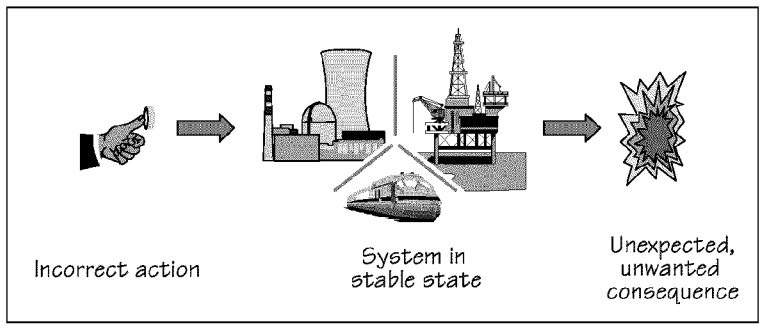

Figure 1: A simple accident model

2.2 Intermediate Accident Model

The simple accident model was gradually extended to recognise both the contribution of latent system states, and the complexity of conditions that could lead to an incorrectly performed human action, cf. Figure 2 – eventually ending by the extreme notion of "error forcing" conditions (Cooper et al., 1996). The complexity of working conditions relaxed the strong assumption of "human error mechanisms", and encouraged descriptions of how human actions were affected by the conditions under which they took place. The latent system conditions – originally called latent system failures (Reason, 1992) – can be precarious conditions brought about by unsound practices of work, as well as consequences of earlier failures. As the name implies, the latent conditions remain undetected until changed circumstances turn them into manifest failures that require rapid responses – usually on top of other events that demand attention. Latent system conditions in safety functions are particularly malicious, because they decrease the safety level without anybody knowing about it while the process is running. In addition, when the safety system is needed, the lack of appropriate responses may lead to a temporary or permanent loss of control of the situation.

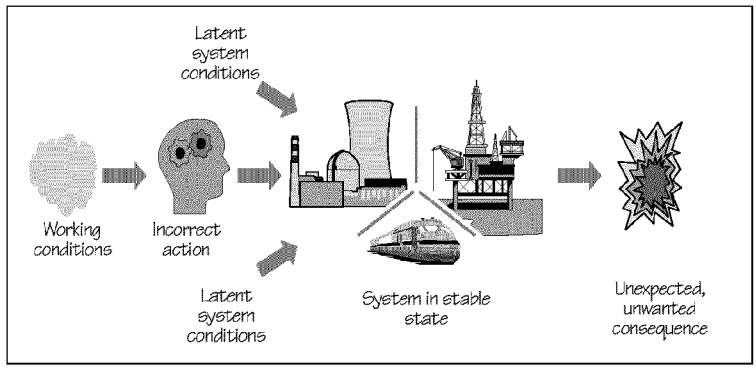

Figure 2: Intermediate accident model

2.3 Contemporary Accident Model

The common approach to analysing and understanding accidents has in the 1990s further shifted the perspective from individual actors to the organisational context. Although the actions – and failures – of individuals still constitute the initiating event, it is necessary to understand the complexity of the working environment, not least the existence of latent conditions. An excellent account of this work has been provided by Reason (1997), which emphasises the concept of organisational safety and how defences may fail.

In the current approach, as shown in Figure 3, the immediate or proximal cause of the accident is a failure of people at the sharp end who are directly involved in the regulation of the process or in the interaction with the technology (Reason, 1990; Woods et al., 1994). A combination of factors that relate to either the human, the technological or the organisational parts of the system – the so-called Man-Technology-Organisation or MTO perspective – is used to explain this failure. The failure at the sharp end is, however, only the triggering condition. The accident does not occur unless there is also a number of latent conditions that suddenly become "active". Furthermore, the outcomes of the failure at the sharp end are both overt and hidden consequences, the latter possibly becoming latent conditions that during a future event may affect the safety of the system.

In addition to the immediate cause, this view also assumes a set of background or proximal causes that are due to function failures at the blunt end. People at the blunt end are to a large extent responsible for the conditions to which by people at the sharp end are exposed, but are themselves isolated from the actual operation. They can be managers, designers, regulators, analysts, system architects, instrument providers, etc. It is the ambition of the contemporary perspective to account for the complex interactions of distal and proximal causes, as well as for the temporal relations, i.e., the way in which past, present, and future are coupled.

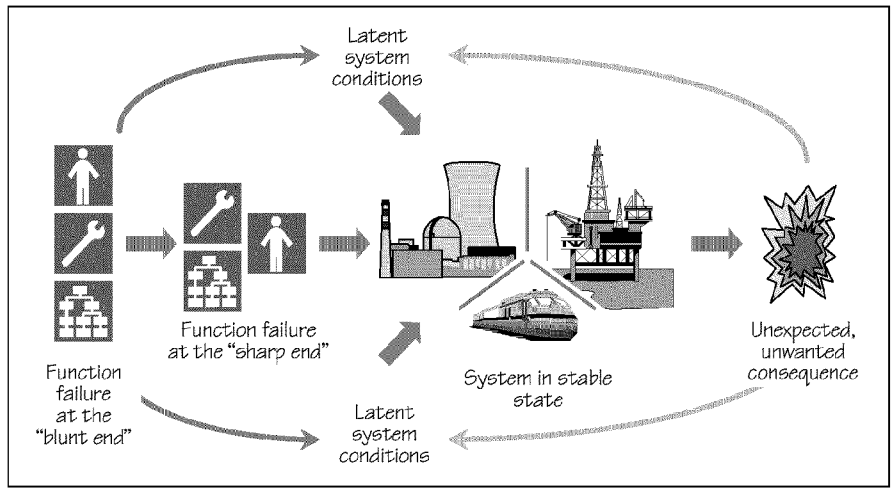

Figure 3: A contemporary accident model

2.4 The Nature Of Causes

Despite these developments, specifically the increasing sophistication in accounting for the organisational determinants of accidents, there is an almost intransigent preference to refer to "human error" as a singular concept. This preference persists in spite of the clear demonstration from the history of accident analysis that the notion of a cause itself is an oversimplification. As pointed out by Woods et al. (1994), a cause is an attribution after the fact or a judgement in hindsight, rather than an objective, unequivocal fact. The determination of the "cause" is a relative rather than absolute process, hence pragmatic and social rather than scientific and deductive. According to this view, a cause can be defined as **the identification, after the fact, of a limited set of aspects of the situation that are seen as the necessary and sufficient conditions for the effect(s) to have occurred**. A cause is in general acceptable:

- If it can unequivocally be associated with a system structure or function (people, components, procedures, etc.).

- If it is possible to do something to reduce or eliminate the cause within accepted limits of cost and time.
- If it conforms to the current "norms" for explanations.

This acknowledgement notwithstanding, accident models are firmly entrenched both in the idea that a "true" or root cause can be found, and in the idea that "human errors" necessarily must be part of the explanations. The result is that accident models become oversimplified, as shown by the left side of Figure 4. According to this view, the accident is first characterised in terms of the external error mode. Next, a suitable cause is expressed as a combination of likely psychological "error mechanisms" and performance shaping factors, where the latter can only exert their influence through the former. The contrasting view, shown by the right side of Figure 4, is consistent with the principles of cognitive systems engineering (Hollnagel, 1998a; Woods, et a., 1994). The search for causes necessarily begins in the same manner by the external error mode or manifestation of the performance failure. But rather than assuming that the proximal cause must necessarily involve a "human error mechanism", or indeed even be attributable to an individual, the search considers the context of the socio-technical system as a whole. In the remaining part of this paper I will argue that this difference in perspectives has significant consequences for how performance predictions are made and how failures are anticipated.

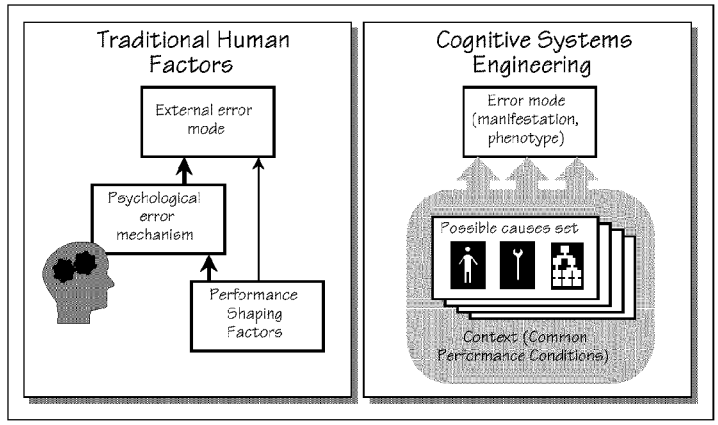

Figure 4: Two perspectives on causation

3. APPROACHES TO PERFORMANCE PREDICTION

As mentioned in the introduction, performance prediction has on the whole been separated from accident analysis. The reality of this separation becomes obvious if one tries to apply any of the established "human error models" for prediction. Indeed, neither the methods nor the categories used allow an easy reversal of the direction of going into the past to going into the future.

One reason for this is that accident models focus on **error types**, where as performance prediction must focus on **error modes**. An error type is a category that is based on and derives its meaning from an underlying model of human action – specifically of human information processing or "cognition in the mind". Well-known examples of model-defined error types range from errors of omission and commission to skill-based, rule-based, and knowledge-based lapses and mistakes. An error type is linked to a specific model of the processes behind human action, and to how these processes mail fail. In contrast to that, an error mode – or human failure mode – is a category that refers to a description of observable manifestations. Error modes may even be logically defined by referring to the small number of physically possible failures (Hollnagel, 1993b). Error modes thus refer to aspects of duration, time, direction, force, sequence, etc. As an example, performing an action too late is an error mode that refers to the aspect of timing of an action. "Action too late" is an overt manifestation and does not in itself make any assumptions of what lay behind it. Performing an action too late may also be described in the language of error types, i.e., referring to what the underlying cause was. Depending on the theoretical stance of the analyst, it may be described as an error of commission or as a rule-based mistake. For accident analysis it may be important to construct an acceptable explanation of the conditions and causes that lead to an accident, hence to focus on error types. For performance prediction it is

important to identify the types of incorrect actions that can occur, regardless of what the causes may be, hence to focus on error types.

Performance prediction has traditionally been pursued as a separate activity in the form of Human Reliability Assessment (HRA), which is the established way of finding the human failure probabilities required by Probabilistic Risk Assessment (Kirwan, 1994). Performance prediction is, however, also an integral part of system design. During the design process choices are made which involve assumptions about the responses of humans and technical systems in future situations. As commonly practised, system design has two major objectives. The first is to ensure that the system performs as required, i.e., that it meets the functional requirements. The second is to avoid that unexpected events happen and that failures occur. The former usually takes up the major part of the design process, whereas the latter is treated more sporadically – almost as a stepchild.

The concern for system failures has grown significantly over the last decades almost as a realisation that system failures generally are unavoidable (Perrow, 1984). The anticipation of system failures is guided by the dominating scientific paradigm, which traditionally is one of decomposition – in particular the decomposition of a system into its "natural" parts, humans and machines. This paradigm has been firmly established by disciplines such as human factors (ergonomics) and human-computer interaction. Since the reliability of modern technology is quite high, the logic of the decomposition approach has forced the focus onto issue of human reliability, usually as single individuals and more rarely as groups or organisations.

3.1 HRA And Human Performance Failure

Due to the influence of accident analysis and HRA, the common approaches to performance predictions have focused human performance – or rather, human performance failures. Performance prediction, as practised by HRA, confines itself to an investigation of the ways in which actions can possibly fail, often referred to as action error modes – or just error modes. In doing so, the likelihood of failure is seen as an attribute of human actions *per se*, often expressed in terms of a "human error probability" (HEP). This is quite consistent with the information processing view, where specific internal "error mechanisms" are assumed to exist. If a function can be seen as an attribute of a component, it follows that the possibility of function failure can be considered for the component by itself, although it is acknowledged that the circumstances or context may have some influence. In HRA the circumstances have been encapsulated by the set of performance shaping factors, which exert their influence in a simple, additive fashion. Yet the likelihood of a component function failure – read: "human error" – is calculated or assessed prior to, hence independent of, the effects of the performance shaping factors.

Anticipating failures of joint human-machine systems requires an underlying model. This should not be a model of human information processing in disguise, but a model of how human performance is determined by – hence reflects – the context or circumstances, i.e., a model of joint system performance rather than of human actions. This type of model corresponds to the notions of distributed or embedded cognition (Hutchins, 1995), although neither of these have been used to consider performance prediction specifically. A concrete expression of these ideas is found in the contextual control models (Hollnagel, 1998b), which describes how humans and technology function as joint systems, rather than how humans interact with machines. The contextual control models emphasise how human-machine co-operation maintains an equilibrium rather than how human-computer interaction can be optimised. The emphasis is thus on "cognition in the world" rather than "cognition in the mind".

3.2 "Human Error" As Noise Or Signal

It is assumed both by HRA and accident analysis that it is reasonable to consider the inherent variability of human performance by itself, specifically that a performance failure is an attribute of the "human component" rather than of the circumstances during which actions take place. In this sense the "human error" is –

metaphorically, et least – the signal rather than the noise. This assumption is strangely inconsistent with one of the main tenets of the information processing approach, which states that:

> "A man, viewed as a behaving system, is quite simple. The apparent complexity of his behavior over time is largely a reflection of the complexity of the environment in which he finds himself."
> (Simon, 1972, p. 25)

If this assumption was used as the basis for anticipating failures, then the focus would be on the variability of the environment or circumstances and not on the possibility of a failure of the "human component". Or rather, the possibility of failure would be an attribute of the context and not of the human. More recently, a similar notion has been expressed specifically addressing the issue of error management:

> "The evidence from a large number of accident inquiries indicates that bad events are more often the result of error-prone situations and error-prone activities, than they are of error-prone people."
> (Reason, 1997, p. 104)

Interestingly enough, a number of HRA methods can be seen as supporting this view. The classical principle of time-reliability correlation (TRC, cf. Hall et al., 1982) is an expression of the idea that the likelihood of failing in performing an activity is a function of time – although in this case it is time after the onset of an accident rather than time available. A more sophisticated version of the same principle is found in the notion of "error forcing conditions", although a determining factor here is time available rather than elapsed time (Cooper et al., 1996). The sophistication is due both to the set of conditions that may "force" an error, and the more detailed description of possible error modes. The common feature is that the possibility of performance failure is an attribute of the conditions rather than of the humans.

A closer inspection of a commonly used HRA method such as HEART (Williams, 1988) also reveals the dominance of the circumstances over the individual. Firstly, HEART only refers to the possible failure of an action, but no to specific failure types. Secondly, the characterisation is related to different tasks, which actually means different task conditions. This can be substantiated by a gentle reinterpretation of the basic HEART table, as shown in Table 1.

Table 1: Description of failure types and causes in HEART

Generic tasks	Context or set of circumstances
A. Totally unfamiliar, performed at speed with no idea of likely consequence.	High time pressure, unfamiliar situation
B. Shift or restore system to a new or original state on a single attempt without supervision or procedures	Lack of supervision and procedures
C. Complex tasks requiring high level of comprehension and skill	High task complexity
D. Fairly simple task performed rapidly or given scant attention.	Simple tasks of limited significance
E. Routine, highly-practised, rapid task involving relatively low level of skill.	Routine or highly familiar tasks
F. Restore or shift system to original or new state following procedures, with some checking .	Following a procedure
G. Completely familiar, well-designed, highly practised routine task, oft-repeated and performed by well-motivated, highly trained individual with time to correct failures but without significant job aids.	High-routine task with no time pressure
H. Respond correctly to system event when there is an augmented or automated supervisory system providing accurate interpretation of system state.	Task with monitoring and highly supportive MMI
M. Miscellaneous tasks for which no description can be found.	No specific characteristics

Even allowing for the limited objectivity of the reinterpretation, it is a demonstrable fact that the major source of variability, which determines the likelihood of a failure, is ascribed to the context or circumstances. In other words, the specific working conditions are the signal while the individual human error probability is the noise. The possibility of performance failure is thus an attribute of the conditions rather than of the humans.

3.3 Look To The Performance Conditions

The consequence of this line of argument is that the variability of human performance constitutes the noise rather than the signal. Conversely, the main determinant of performance quality – and specifically of performance failure – comes from the context or the circumstances. The possibility of failure is therefore an attribute of the joint system rather than of any of its components. It follows from this that the anticipation of system failures should concentrate on developing effective ways of describing how joint system performance depends on the conditions rather than on the potential for human failures.

Specifically, predictions should be about how the joint system can lose control of the situation, rather than about whether the human will make a single failure. This would also acknowledge the fact that a human failure is just a single event that requires other conditions to result in an accident – cf. the extended accident model. A practical implementation of this principle can be found in the basic method for performance prediction that is part of CREAM (Hollnagel, 1998a). Here an assessment of the common performance conditions leads to an overall prediction of how likely the operator, hence the joint system, is to lose control. This is prediction is made without considering the failure probability for specific actions, or even describing the tasks at the level of component actions.

The CREAM approach assumes that the likelihood of a failure (the proverbial "human error") depends on the working conditions rather than on the propensity of humans to screw up. It also assumes that it are the system functions and (latent) conditions that determine whether an action failure turns into an accident. In CREAM the basic method for performance prediction starts by characterising the context in terms of a small number of Common Performance Conditions (CPC). Unlike the traditional HRA approaches, the CPCs are assumed to depend on each other in a manner shown by the model shown in Figure 5. In the basic application of the method, the effect of the CPCs alone is sufficient to provide an overall characterisation of the situation, hence of the likelihood of losing control. In the more detailed use, specific error modes – and their probabilities – may be determined, but only on the background of the expected performance conditions. A detailed account of this approach is provided in Hollnagel (1998a).

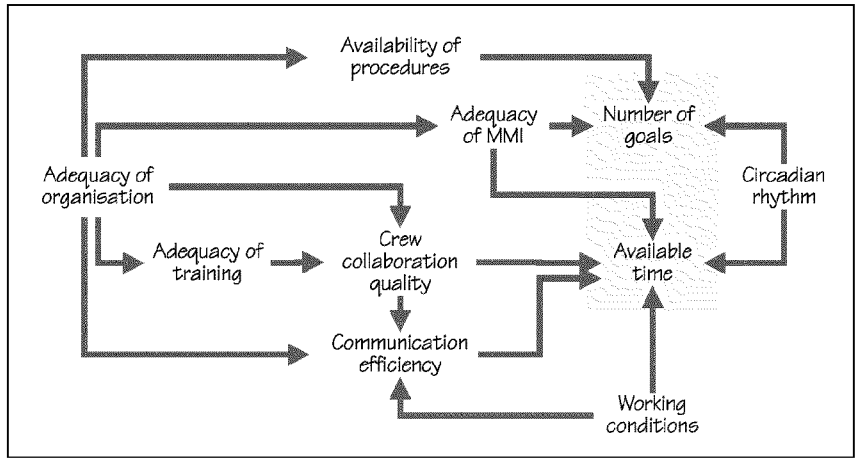

Figure 5: The CPC dependency model in CREAM

4. CONCLUSIONS

I have in this paper tried to argue that there is a need to understand the complexity of joint system performance, and the dynamics of human-machine interaction. Instead of focusing on discrete actions taking

place at single points in time, one should focus on how actions develop over time – on how an event unfolds and how the joint system strives to maintain an equilibrium. It is during this dynamic process that prior events have future consequences, depending on how the conditions change.

A concrete criticism against the established practice is that the repertoire of methods for performance prediction and anticipating failures do not fully reflect the lessons from accident analysis. Despite an impressive and growing amount of evidence, performance prediction remains focused on the notion of "human error". As argued elsewhere (Hollnagel, 1998a), the concept of "human error" is an artefact of the models and methods that have been used, leads to a view of performance failure as an attribute of the individual – and of human cognition – rather than as an attribute of the context. In order to overcome this bias, we need to develop better models both of how performance conditions affect the likelihood of failure or losing control and of how coincidences can occur and barriers fail. This might also relieve us from hunting after the elusive human error probability and the impossible task of controlling the conditions of observation, and instead look to what is really important – the natural contexts in which people have to work.

5. REFERENCES

Cojazzi, G. & Pinola, L. (1994). *Root cause analysis methodologies: Trends and needs.* In G. E. Apostolakis & J. S. Wu (Eds.), Proceedings of PSAM-II, San Diego, CA, March 20-25, 1994.

Cojazzi, G. (1993). *Root cause analysis methodologies. Selection criteria and preliminary evaluation* (ISEI/IE/2442/93). JRC Ispra, Italy: Institute for Systems Engineering and Informatics.

Cooper, S. E., Ramey-Smith, A. M., Wreathall, J., Parry, G. W., Bley, D. C., Luckas, W. J., Taylor, J. H. & Barriere, M. T. (1996). *A technique for human error analysis (ATHEANA)* (NUREG/CR-6350). Washington, DC: US Nuclear Regulatory Commission.

Dougherty, E. M. Jr., & Fragola, J. R. (1988). *Human reliability analysis. A systems engineering approach with nuclear power plant applications.* New York: John Wiley & Sons.

Hall, R. E., Fragola, J. R. & Wreathall, J. (1982). *Post-event human decision errors: Operator action trees/time reliability correlation* (NUREG/CR-3010). Washington, DC.: USNRC.

Hollnagel, E. (1993a). *Human reliability analysis: Context and control.* London: Academic Press.

Hollnagel, E. (1993b). The phenotype of erroneous actions. *International Journal of Man-Machine Studies, 39,* 1-32.

Hollnagel, E. (1998a). *Cognitive reliability and error analysis method – CREAM.* Oxford: Elsevier Science.

Hollnagel, E. (1998b). Context, cognition, and control. In Y. Waern, (Ed.). *Co-operation in process management - Cognition and information technology.* London: Taylor & Francis

Hutchins, E. (1995). *Cognition in the wild.* Cambridge, MA: MIT Press.

Kirwan, B. (1994). *A guide to practical human reliability assessment.* London: Taylor & Francis.

Park, K. S. (1987). *Human reliability. Analysis, prediction, and prevention of human errors.* Amsterdam: Elsevier.

Perrow, C. (1984). *Normal accidents: Living with High-Risk Technologies.* New York: Basic Books.

Reason, J. T. (1990). *Human error.* Cambridge, U.K.: Cambridge University Press.

Reason, J. T. (1992). The identification of latent organisational failures in complex systems. In J. A. Wise, V. D. Hopkin & P. Stager (Eds.), *Verification and validation of complex systems: Human factors issues.* Berlin: Springer Verlag.

Reason, J. T. (1997). *Managing the risks of organizational accidents.* Aldershot, UK: Ashgate.

Senders, J. W. & Moray, N. P. (1991). *Human error. Cause, prediction, and reduction.* Hillsdale, NJ.: Lawrence Erlbaum.

Simon, H. A. (1972). *The sciences of the artificial.* Cambridge, MA.: The M. I. T. Press.

Williams, J. C. (1988). *A data-based method for assessing and reducing human error to improve operational performance.* Proceedings of IEEE 4th Conference on Human factors in Power Plants, Monterey, CA, 6-9 June.

Woods, D. D., Johannesen, L. J., Cook, R. I. & Sarter, N. B. (1994). *Behind human error: Cognitive systems, computers and hindsight.* Columbus, Ohio: CSERIAC.